An Attempt to Explore the Similarity of Two Proteins by Their Surface Shapes

Xiang-Sun Zhang Zhong-Wei Zhan Yong Wang
Ling-Yun Wu

Academy of Mathematics and Systems Science, CAS, Beijing 100080, China

Abstract *Protein structure comparison is one of the most important problems in Bioinformatics. In this paper, we develop a convex hull expression of a protein and propose a grid method to evaluate similarity between two convex hulls. The purpose of introducing the convex hull expression is to investigate the relationship between the surface shape similarity and the protein function similarity. Extensive numerical experiments shows that proteins in a family do have surface shape similarity and proteins in different families have their shapes in quite large difference. To explore the shape similarity of proteins in different families, we observe the possible local shape similarity by reorganizing the convex hull's supporting plane information into a set of local shape information. And comparing the local shape information of two proteins leads us to find some local shape similarity for some pairs of proteins whose structure similarity is difficult to evaluate by traditional comparison methods. The biological meaning of local shape similarity remains unclear.*

Keywords Protein Structure Similarity, Protein Structure Comparison, Convex Hull, Grid Algorithm, Vector Sets Comparison

1 Introduction

It is well known that the elucidation of protein function is one of the key tasks of molecular biology. How protein functions and interacts with other molecules is closely related to their structure information. Therefore protein structure comparison has been one of the most important problems in Bioinformatics. The purpose of researchers is to find the evolutionary connections of proteins by comparing their 3D structures.

The tertiary structures of proteins have been experimentally recorded by X-ray crystallography or NMR spectroscopy and deposited in the database PDB [1]. There are total 31,059 structures of proteins, nucleic acids, protein-nucleic acid complexes and carbohydrates in PDB as of May 31, 2005 and it still increases by about 30-50 entries every week. In recent years, a number of different automatic

methods of protein structure comparison have been proposed as indicated in the comprehensive reviews [2, 3, 6, 7, 9, 11, 12]. Most of them are to align proteins' C_α atoms in the backbone. These methods can identify structural resemblances accurately but have some shortages such as long computational time, complex process and large data space.

In this paper, we develop a new idea that focuses on the protein's global shape. Every protein is represented by its convex hull and the protein surface is approximated by a set of facets, $f_i, i = 1, \cdots, m$, of the convex hull. The plane on which each facet f_i is is called the supporting plane. We define some measures based on the sets of supporting planes (SPs) instead of the sets of the coordinates of C_α atoms in the traditional protein alignment research to evaluate protein structure similarity.

In section 2, we present out method and numerical results for protein's global shape comparison. Local shape comparison is discussed in section 3. In section 4, we give some conclusion.

2 Method and Results for Protein's Global Shape Comparison

In most existing comparison methods[5, 10], a protein is often viewed as a sequence of C_α atoms described by the position of their centers which is called the backbone of the protein. Suppose that the shape of a protein is completely determined by the backone

$$X = \{x^k\} = \{(x_1^k, x_2^k, x_3^k), k = 1, \cdots, N\} \tag{1}$$

where $(x_1^k, x_2^k, x_3^k) \in R^3$ is the coordinate of a C_α atom.

Let $conv(X)$ be the convex hull of the set X . $x \in conv(X)$ if and only if x can be represented as

$$x = \sum_{j=1}^{k} \lambda_j x^j, \quad \sum_{j=1}^{k} \lambda_j = 1, \quad \lambda_j \geqslant 0, \quad x^j \in X, \quad j = 1, 2, \cdots, k \tag{2}$$

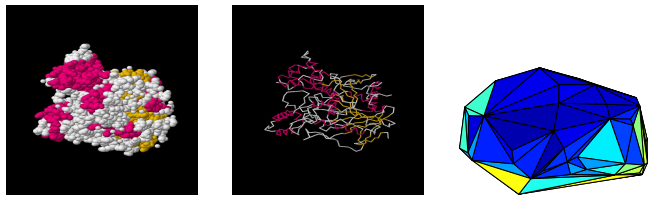
Let $P = \{p_h\}$ be the set of supporting planes (SPs) of $conv(X)$. $P = \{\, p_h | a_h^T x = b_h \,\}$, then $conv(X)$ can be expressed as

$$conv(X) = \{x : a_h^T x \geqslant b_h \quad h = 1, \cdots, H\} \tag{3}$$

where H is the number of SPs.

Every SP has a normal vector, we use them to represent $cov(X)$. Let $V(X)$ be the set of normal vectors of SPs,

$$V(X) = \{\frac{a_h}{\|a_h\|}, h = 1, \cdots, H.\} \tag{4}$$



(a) The space-fill view of 1acj

(b) The backbone view of 1acj

(c) The convex hull of 1acj

Figure 1: A protein and its convex hull

Figure 1 shows the protein 1acj and its convex hull. Figure 1 (a) is the protein graph by Rasmol 2.7.1.1, then only the C_α atoms are cared, and the backbone is shown in (b). Through computation, its convex hull is shown in (c).

Remark 1: One of the advantages of the convex hull representation lies in the effect of data compression. Suppose that every three C_α atoms of a protein possibly form a supporting plane. Then for a protein with $N = n^3$ amino acid residues, there are approximately totally N^3 plans, but most of them can not be supporting plans. If a protein is folded in a cubic with edge length n, then the total number of plans constituted by the atoms at the surface is upper bounded by $O(N^2)$. In fact, for each atom, only a constant number, which is strongly less N, of supporting plans are passing through it. That is, we expect that the total number of supporting plans in the convex hull is $O(N)$ and with a small constant c, i.e. cN. The numerical experiments have shown the validity of this estimate. For example, we examine the protein 1acj in Figure 1. There are totally 4095 atoms and 528 C_α atoms. But when considering the convex hull expression, there are only 120 SPs. So the protein data is reduced to an sequence with rather short space. The contrast between the protein length and the number of faces of a protein's convex hull for a bundle of proteins is shown in Figure 2, from which we can see that the new representation indeed reduces the data input for comparison and the tendency is more clear when the protein size gets larger.

Remark 2: The information of SPs is unordered not like the set of coordinates of the C_α atoms. Lack of order information implies both advantages and disadvantages. In one hand, we need not care about the relative order of elements in the comparison process. This can decrease the computational time. But on the other hand, it could lead to non-significant comparison results.

We propose a grid method to compute the fitness degree between two SP sets to measure protein structure similarity. Consider two proteins labeled A and B($H^A > H^B$). A pairwise shape comparison problem can be formulated as the following mixed integer programming:

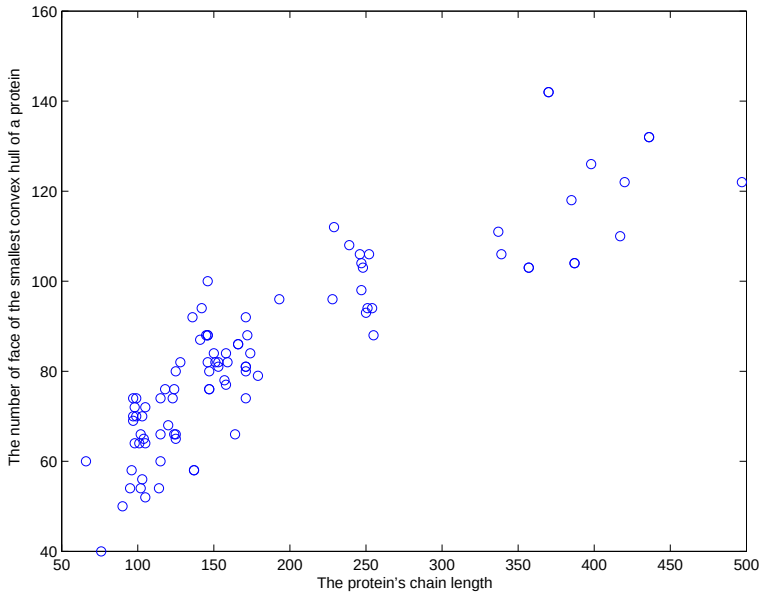


Figure 2: Relationship between the protein length and the number of faces of the convex hull of a protein

$$minimize \qquad E(S,R) = \sum_{i=1}^{H^A} \sum_{j=1}^{H^B} g_{ij} \ \|a_i^A - Ra_j^B\| \quad for\ G, R \tag{5}$$

$$subject\ to \qquad \sum_{j=1}^{H^B} g_{ij} = 1,\ i = 1, \cdots, H^A \tag{6}$$

where $a_i^A \in V^A$, $a_j^B \in V^B$, V^A and V^B represent the sets of normal vectors of protein A and protein B, respectively. $R \in R^{3\times 3}$ is a rotation matrix, and binary entries g_{ij} describe matching of V^A and V^B:

$$g_{ij} = \begin{cases} 1 & \text{if vector-i in } V^A \text{ matches vector-j in } V^B \\ 0 & otherwise \end{cases} \tag{7}$$

In the definition of the measure function $E(S, R)$, we set the protein with more SPs, say A, as target and rotate the protein with fewer SPs to match it. Let S_i denote the area of f_i. Instead of solving the difficult mixed integer programming we redefine a measure function without the matching variable g_{ij}. We confine an a_i only matching a_js with S_j almost equal to S_i, then the new measure function is

Protein Name	Number of SPs	d_{AB}	N^G	Protein Name	Number of SPs	d_{AB}	N^G
1hlm00	84	—	—	1mba00	100	0.4521	0
1hbg00	78	0.4938	2	1ash00	80	0.4599	1
1gdj00	78	0.4853	1	1ithA0	87	0.4456	1
1babA0	82	0.4617	3	1vhbA0	72	0.4970	0
1flp00	94	0.4789	2	1ew6A0	88	0.4552	0
1lhs00	77	0.4833	1	1dlwA0	79	0.4492	8
1eca00	92	0.4588	4	1h97A0	80	0.4829	4
1sctA0	85	0.4436	1	1kr7A0	74	0.4872	4

Table 1: Globins family (protein 1hlm00 is the family representative according to CATH database, the measure numbers in the third, fourth, seventh and eighth column are from the comparison of the proteins in the family to the representative.)

Protein Name	Number of SPs	d_{AB}	N^G	Protein Name	Number of SPs	d_{AB}	N^G
1ra900	82	—	—	1vdrB0	82	0.4131	0
3dfr00	84	0.4713	2	1df7A0	70	0.4491	0
8dfr00	86	0.4610	0	1cz3A0	88	0.4505	1
1aoeA0	83	0.4554	0	————			

Table 2: Dihydrofolate reductases family (family representative:1ra900)

$$E = \sum_{i=1}^{H^A} \min_{\{j:\ |S_j - S_i| < \gamma S_i\}} (\|a_i^A - Ra_j^B\|) \tag{8}$$

where $\gamma = 0.2$.

The similarity of two structures is evaluated by the average distance of matched SP:

$$d_{AB} = \frac{E}{H^A - N^G} \tag{9}$$

$$N^G = |i : \{j : |S_j - S_i| < \gamma S_i\} = \emptyset| \tag{10}$$

where N^G represent the number of gaps.

To solve eqns.(8), this paper adopts a grid method that can get results in 5 minutes of computer time on a PIII CPU. Numerical results indicate that with the shape measure (9), the proteins in the same family are shape-similar. The data is organized as follows. We set the representative protein of a family as target, compare other proteins in this family to it as shown in Table1, Table2 and Table3. The measure between them is generally less than 0.5 . On the other hand, comparing the proteins belonging to different families will lead to large measures and many gaps (Table4, Table5). It means that proteins in different families are quite different from the global shape view. According to Table4 and Table5, we can also find that if two proteins are very similar in structure, the grid method get very small measure numbers (1colA0–1colB0, 1thv00–1thu00).

Protein Name	Number of SPs	d_{AB}	N^G	Protein Name	Number of SPs	d_{AB}	N^G
1rcy00	74	—	—	1aozA3	88	0.4690	6
1jer00	64	0.4207	1	1cyx00	95	0.4617	2
1plc00	76	0.4560	4	1occB2	74	0.4717	1
1paz00	76	0.4738	0	1a65A2	73	0.4952	3
1jzeA0	76	0.4569	3	1a65A3	74	0.4945	4
1aozA1	62	0.4745	4	2cuaA0	68	0.4902	0
1aozA2	76	0.4996	5	1kcw01	70	0.4830	7
1nif01	72	0.4823	3	1kcw02	60	0.4944	1
1nif02	71	0.4879	5	1qniA2	70	0.3936	2
2cbp00	60	0.4246	4	1ehkB0	84	0.4919	5

Table 3: Cupredoxins family (family representative:1rcy00)

Protein Name	Number of SPs	CATH code	d_{AB}	N^G
1colA0	110	1.10.490.30.1.1.1	—	—
proteins belonging to same family				
1colB0	109	1.10.490.30.1.1.1	0.1371	0
1cii03	98	1.10.490.30.2.1.1	0.3902	2
1ddt02	93	1.10.490.40.1.1.1	0.4496	0
1hlm00	84	1.10.490.10.1.1.1	0.4800	0
1cpcA0	66	1.10.490.20.1.1.1	0.4981	0
proteins belonging to different family				
1cuk03	46	1.10.8.10.2.1.1	0.5198	9
1vdfA0	48	1.20.5.10.1.1.1	0.8891	2
1ahl00	36	2.20.20.10.1.1.1	0.6616	16
1pdc00	46	2.10.10.10.1.1.1	0.5927	3

Table 4: Protein 1colA0 (belonging to α class in the CATH database[8]) is compared with other proteins from the global shape view

Protein Name	Number of SPs	CATH code	d_{AB}	N^G
1thv00	90	2.60.110.10.1.1.1	—	—
proteins belonging to same family				
1thu00	96	2.60.110.10.1.1.1	0.2385	0
1du5A0	98	2.60.110.10.1.3.1	0.3723	1
1aun00	89	2.60.110.10.1.2.1	0.2922	0
proteins belonging to different family				
1d00A0	72	2.60.210.10.1.1.1	0.5009	1
1cauB0	79	2.60.120.10.1.1.1	0.5983	0
1npoA0	58	2.60.9.10.1.1.1	0.5728	5
1ahl00	36	2.20.20.10.1.1.1	0.7060	5
1pdc00	46	2.10.10.10.1.1.1	0.5794	14
1cuk03	46	1.10.8.10.2.1.1	0.5138	16
1vdfA0	48	1.20.5.10.1.1.1	0.9629	0

Table 5: Protein 1thv00 (belonging to β class in the CATH database) is compared with other proteins from the global shape view

3 Local Shape Comparison

Global shape comparison show that proteins in different families may have big difference in overall shapes. But it does not reject possible local resemblance of the two proteins with long distance in the classification tree. The motivation to explore the local resemblance is due to a conjecture that sometimes the function of a protein is decided by its local shape but not the global shape, such as a docking part of a protein.

To consider the local similarity we reorganize the set of supporting planes into subsets to represent the local information of a protein surface. Consider a protein backbone X, let $P = \{p_h\}$ and $V = \{a_h\}$ denote the set of SPs and the set of normal vectors respectively. We define a local SP Set (LSP) with respect to an SP p_h as follows:

$$\widetilde{P_h} = \{p_i \in P |\ \exists x \in X, x \in p_i, x \in p_h\} \tag{11}$$

where p_h is called the featured plane of the LSP $\widetilde{P_h}$ and let $A(p_n) = \{a_i : P_i \in \widetilde{P_h}\}$ be the neighboring set of a_h. (Similarly, facet f_h is called a featured facet).

From the above definitions, we can see that a LSP is composed of a featured plane and all of its adjacent SPs. In this way, we focus the comparison area from the total surface of convex hull into some local areas and compare the LSPs to find the local resemblance in protein structures. The following objective function $\widetilde{E}(R)$ count for the total number of matched local sets while $E_{ij}(R)$ measures the similarity degree of two subsets.

$$\widetilde{E}(R) = -\sum_{i=1}^{\widetilde{n}} f(\min_{j|\ |S_j - S_i| < \gamma S_i} E_{ij}(R) - \varepsilon) \tag{12}$$

$$E_{ij}(R) = (\sum_{b_i \in A(p_i)} \min_{b_j \in A(p_j)} \|b_i - Rb_j\|) + \theta \|a_i - Ra_j| \tag{13}$$

where $A(p_i), A(p_j)$ represent the neighboring sets of a_i and a_j respectively. We suppose $|A(p_i)| \leqslant |A(p_j)|$. S_i, S_j represent the area of LSP-i and LSP-j's featured facet respectively. $f(x)$ is defined as

$$f(x) = \begin{cases} 1 & x < 0, \\ 0 & x \geqslant 0. \end{cases} \tag{14}$$

The LSPs are ordered by their featured facet's area. Because the LSPs of a convex hull are intersectant, we set $\widetilde{n} = \lfloor 0.8H^A \rfloor$. It means that we only consider those LSPs with large featured facet. For the other three parameters γ, θ and ε, we set

$$\gamma = 0.2,\ \theta = 5,\ \varepsilon = 1.6, \tag{15}$$

Based on this new measure function, we compare 12 pairs of proteins, which have been reported[4] are difficult to compare by traditional method. The results are summarized in Table 6. Obviously, these protein pairs have structural resemblances

Protein Pairs	DIFF.	Number of SPs	Number of Matched LSPs	Average Distance of Matched LSPs
1rcb–2gmfA	12.7	68–80	3	0.1142
2afnA–1aozA	14.6	64–64	2	0.0983
1ubq00–1fxiA0	15.3	60-66	3	0.1209
1bgeB0–2gmfA0	15.4	76–80	1	0.0736
3hlaB0–2rhe00	16.4	54–57	2	0.0837
3chy00–2fox00	17.3	82–84	2	0.1278
2azaA0–1paz00	18.0	76–76	2	0.1133
1molA0–1cew00	18.1	64–66	3	0.1068
1ten00–3hhrB2	20.0	64–58	1	0.0868
2trxA0–1gp1A0	20.0	76–86	2	0.1009
1tie00–4fgf00	20.0	61–74	2	0.1345
2rhe00–1cid01	20.0	57–74	1	0.0988

Table 6: Comparisons of some difficult protein pairs using the number of matched LSPs

in at least one LSP. It means that they are very similar in some surfaces of the convex hulls and the improved model can help us find the local similarities of proteins. But we have not displayed the biological meaning of the local shape resemblance between these proteins in this paper.

4 Conclusion

In this paper, a novel method is proposed to explore the protein structure comparison problem from the view of the protein's shape. It is indicated that such a measure of protein similarity by the surface similarity is capable of providing insight into a protein's 3D structure and capturing some factors to assess protein structure similarity in an automatic way.

From the view of implementation, our new approach to measure structure difference between proteins have two main advantages. First, our method can be implemented automatically. Second, it is easy to implement and very fast as described in the previous sections. Numerical results indicate that the new model has good performance for the proteins from the same family. Furthermore, by comparing the LSPs of proteins, we can find the local similarities in some surfaces which are difficult to be identified in global comparison.

From the numerical results, we can see that the shape of protein is an important characteristic within a protein family. But in the class and fold level, it has less influence. To effectively classify proteins, we maybe need to integrate the shape information into other shape indices. This is our future research direction. In fact the established automatic protein classification tools use large number of indices that describe protein's different profiles. From this point of view, the protein surface shape along has displayed its potential in protein classification, then is worth to study further.

References

[1] H. M. Berman, J. Westbrook and Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne. The protein data bank. *Nucl. Acid Res.*, 2000, 28:235–242.

[2] Oliviero Carugo and Sandor Pongor. Recent progress in protein 3d structure comparison. *Curr. Protein Pept. Sci.*, 2002, 3(4):441–449.

[3] I. Eidhammer, I. Jonassen and W. R. Taylor. Structure comparison and structure patterns. *Journal of Computational Biology*, 2000, 7:685–716.

[4] D. Fischer, A. Elofsson, D. Rice and D. Eisenberg. Assessing the performance of fold recognition methods by means of a comprehensive benchmark. *Pac. Symp. Biocomput.*, 1996, 300C318.

[5] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 1993, 233:123–138.

[6] Liisa Holm and Chris Sander. Mapping the protein universe. *Science*, 1996, 273(2):595–602.

[7] T. Kawabata and K. Nishikawa. Protein structure comparison using the markov transition model of evolution. *Proteins: Structure, Function and Genetics*, 2000, 41:108–122.

[8] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells and J. M. Thornton. CATH – A Hierarchic Classification of Protein Domain Structures. *Structure*, 1997, 5(8):1093–1108.

[9] Joanna M. Sasin, Michal A. Kurowski and Janusz M. Bujnicki. STRUCLA: A WWW meta-server for protein structure comparison and evolutionary classification. *Bioinformatics*, 2003, 19 Suppl:1252–1254.

[10] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 1998, 11:739–747.

[11] A. P. Singh and D. L. Brutlag. Protein structure alignment: a comparison of methods. Technical report, Stanford University, CA, 1999.

[12] Yusu Wang. Pairwise protein structure comparison techniques. Technical report, Center for geometric Computing, Department of Computer Science, Duke University, NC, 2002.